Application of Principal-Component Analysis on Near-Infrared Spectroscopic Data of Vegetable Oils for Their Classification

Tetsuo Sato

Department of Crop Breeding, Kyushu National Agricultural Experiment Station, Ministry of Agriculture, Forestry and Fisheries (MAFF), Nishigoshi, Kumamoto-ken, Japan 861-11

In the near-infrared (NIR) spectra of oil, information about fatty acid composition is concentrated in the range of 1600–2200 nm. Principal-component analysis (PCA) was applied on the standardized full NIR spectral data of this region for vegetable oils to totally capture the NIR spectral pattern. Nine varieties of vegetable oils (soybean, corn, cottonseed, olive, rice bran, peanut, rapeseed, sesame and coconut oil) could be successfully classified from their PCA scores. Examining the contribution of wavelengths to PCA scores showed that wavelengths with a high loading weight were assigned to characteristic absorption regions that correspond to specific fatty acid moieties. This classification is related to the fatty acid composition of an oil, and it can be carried out rapidly and easily after eigenvectors were obtained.

KEY WORDS: Classification, fatty acid composition, near-infrared spectra, principal-component analysis, spectroscopy, vegetable oil.

Recently, near-infrared (NIR) spectroscopy has been recognized as one of the most powerful analytical techniques for determining various constituents in agricultural and food products (1–3). For this purpose, multiple linear regression analysis has been used often for establishing calibration equations prior to routine analyses. However, useful information may be lost in this process because only a few wavelengths are utilized in this statistical method. It is necessary to capture the complete NIR spectral pattern to obtain full information. Further, an attempt has begun to apply NIR data for qualifying or classifying purposes in addition to quantifying purposes. Principal-component analysis (PCA) (4), a data compression method in which little information is lost, has also been applied for analyzing NIR spectral patterns (5–10). However, when PCA is applied, there might be the possibility that samples are classified according to their level of moisture or to their particle size, not to their own peculiar compositional characteristics, because not only absorption characteristics but also scattering determines the pattern of their NIR spectra. As shown in our previous reports (11,12), the NIR spectral pattern of an oil represents its fatty acid composition because NIR absorption bands around 1600–1800 nm and 2100–2200 nm are assigned to straight carbon chains and *cis* double bonds that reflect fatty acid moieties in fat molecules. Because oil consists almost totally of triglycerides and there may be no need to consider either the interference from water absorption or the size effect of scattering particles; the characteristic wavelengths with high loading weight seem to be assigned more easily for oils than for other materials. So, the relationship between the PCA results and the assignments of wavelengths with a high loading weight can be characterized more properly. In this study, PCA was applied to the measured full-range NIR spectral data of some vegetable oils to classify them, and the features of PCA applied to the NIR spectral data were also investigated.

MATERIALS AND METHODS

Samples. Thirty-two samples of nine commercially available vegetable oils were analyzed. They were soybean oil (five kinds), corn oil (four kinds), cottonseed oil (four kinds), olive oil (four kinds), rice bran oil (two kinds), peanut oil (four kinds), rapeseed oil (three kinds), sesame oil (three kinds) and coconut oil (three kinds). The various kinds of oils were purchased from different companies: Eastman Kodak Co. (Rochester, NY); Hayashi-Ichiji Co. (Tokyo, Japan); Kanto Chemicals (Tokyo, Japan); Nakarai Chemicals (Kyoto, Japan); Wako Pure Chemicals (Osaka, Japan) and Yuro Chemicals (Tokyo, Japan).

Chemical measurements. The fatty acid composition of these oils were analyzed by gas chromatography (GC) after methyl esterification (13). The GC analyses of the fatty acid methyl esters were carried out according to a previous report (11).

Physical measurements. A near-infrared spectroscopic instrument (InfraAlyzer 500; Bran + Luebbe, Norderstedt, Germany) was used to measure the NIR transflectance. The spectra were measured twice for each sample. The sample presentation was carried out in a British cup (Bran + Luebbe) as described in a previous report (11,12). The measured wavelengths were from 1600 to 2300 nm in 1-nm increments. Although oils also have strong and sharp NIR absorption bands in the 2300–2400 nm region and weak or no absorption bands around 1100–1600 nm, they have few distinctive features. Further, if these regions were included in the mathematical analysis, they may influence and inhibit the extraction of more useful information from the 1600–1800 and 2100–2200 nm regions. Because information about fatty acid compositions is concentrated in the 1600–2200 nm region (12), it seems better to measure the detailed spectral pattern of this region at 1-nm intervals.

The conditions to calculate the second-derivative spectra were as follows: 2 nm between output points, 2 nm in moving average, 12 nm per derivative segments and 12 nm between derivative segments. Thus, the range of the second-derivative spectrum obtained was 1619 to 2281 nm with a 2-nm interval as obtained with the IDAS software (Bran + Luebbe).

PCA of NIR spectra. As for the raw spectrum, the range of 1600 to 2200 nm with a 2-nm interval (301 data points), and for the second-derivative spectrum, from 1619 to 2201 nm with a 2-nm interval (292 data points) were used in analysis.

Binary files of spectral data were converted to ASCII files by means of the IDAS software to carry out subsequent data processing. Because the levels of the original spectral values were different, the spectra were first standardized as follows: The level of starting point at 1600 (raw spectrum) or 1619 nm (second-derivative spectrum) was set at 0.0, and the maximum or minimum value around 1710–1725 nm was corrected to 1.0 (raw spectrum) or −1.0 (second-derivative spectrum) by the following equation:

$$(A_X)_{stn} = (A_X - A_s)/(A_s - A_m) \qquad [1]$$

where $(A_X)_{stn}$ is the standardized value of A_X, A_X is the original spectral value at X nm, A_s is an original spectral value at the starting point and A_m is a maximum (raw spectrum) or minimum (second-derivative spectrum) of the original spectral value around 1720 nm. Commercially available PCA software, written in BASIC language (14), was rewritten in C language to make it possible to deal with many variables. PCA was carried out with a variance-covariance matrix calculated from the 64 sets of respective raw and second-derivative spectral data from 32 kinds of samples. The average spectrum, standard deviation spectrum and eigenvectors were obtained from this mathematical treatment.

The score plots were also drawn from the PCA scores. The following values were calculated by mean-centering and multiplying with the eigenvector: $\{(A_X)_{stn} - [(A_X)_{stn}]_{av}\} \times \mathrm{E}_i$ where $[(A_X)_{stn}]_{av}$ is the average of the standardized spectral values of oil measured at X nm, and E_i is the eigenvector of the i-th principal component (PC). The sum of these values over the full range of related wavelengths becomes the PCA score of the i-th principal component for each sample.

RESULTS AND DISCUSSION

Fatty acid composition by GC method. Tables 1 and 2 show the average fatty acid composition of each kind of oil tested (nine varieties, 32 kinds). Rapeseed oil also contains a little erucic acid (C22:1), and this might be the reason why its total fatty acid percentage was somewhat less than 100%. It was not detected under the author's analytical conditions (13). However, Tables 1 and 2 show the characteristic fatty acid composition of each oil as follows. Coconut oil consisted of mainly saturated fatty acids, while the others have mainly unsaturated acids. Olive and rapeseed oils contain mostly oleic acid (C18:1). Olive oil also has palmitic acid (C16:0). On the other hand, rapeseed oil contains linoleic acid (C18:2) and linolenic acid (C18:3). Soybean, corn and cottonseed oils contain mainly C18:2, and soybean oil contains a little more C18:3. Further, corn oil contains a little more C18:1, and cottonseed oil has a little more C16:0. Sesame and peanut oils contain almost the same level of C18:1 and C18:2, and peanut oil also has a little more C16:0. Of course, the author measured only a few kinds of vegetable oils, but they had a wide-ranging fatty acid composition, especially the unsaturated moieties (C18:1, C18:2 and C18:3) as shown in Tables 1 and 2.

PCA. Figure 1 shows the average and standard deviation of the standardized raw NIR spectral data of the oil measured, and it also shows the first three eigenvectors of the PCs obtained by PCA treatment. The contribution ratio of these PCs was 88.94, 10.18 and 0.75%, respectively, and their total contribution ratio to the whole variation was 99.87%. The higher PCs were also obtained, but their shapes became less distinct, i.e., they had more fluctuations or noise. The first PC (Fig. 1b) was almost a linearly increasing curve: The spectrum level increased as the wavelength became longer, which is one of the often observed features of the NIR spectrum. It was also similar to the standard deviation spectrum (Fig. 1a). This means that the most astonishing variation in the spectrum was

TABLE 1

Fatty Acid Compositions of Soybean, Corn, Cottonseed, Olive, Rice Bran, Peanut, Rapeseed and Sesame Oils as Methyl Esters[a]

Fatty acid	Soybean (five kinds)	Corn (four kinds)	Cottonseed (four kinds)	Olive (four kinds)	Rice bran (two kinds)	Peanut (four kinds)	Rapeseed (three kinds)	Sesame (three kinds)
C16:0	10.58 ± 0.41 (10.10-11.24)	10.55 ± 0.38 (9.96-10.95)	18.49 ± 0.21 (18.31-18.83)	10.13 ± 0.41 (9.48-10.56)	7.16 ± 2.47 (3.67-9.07)	16.16 ± 0.04 (16.12-16.20)	12.12 ± 0.75 (10.94-13.02)	3.73 ± 0.12 (3.61-3.90)
C16:1	0.07 ± 0.05 (0.03-0.16)	0.12 ± 0.05 (0.07-0.20)	0.71 ± 0.03 (0.67-0.74)	0.99 ± 0.08 (0.90-1.09)	0.13 ± 0.06 (0.05-0.20)	0.10 ± 0.01 (0.08-0.11)	0.24 ± 0.25 (0.06-0.66)	0.25 ± 0.03 (0.21-0.27)
C18:0	3.82 ± 0.23 (3.39-4.04)	1.98 ± 0.09 (1.83-2.08)	2.29 ± 0.07 (2.21-2.40)	3.32 ± 0.33 (3.03-3.85)	4.46 ± 1.99 (1.65-6.04)	1.62 ± 0.01 (1.61-1.63)	2.85 ± 0.18 (2.62-3.04)	1.68 ± 0.06 (1.62-1.77)
C18:1	23.77 ± 0.51 (23.21-24.48)	29.36 ± 3.04 (26.83-34.51)	19.28 ± 1.20 (17.97-21.18)	78.32 ± 1.68 (75.48-79.81)	48.08 ± 7.78 (42.12-59.07)	41.97 ± 0.10 (41.87-42.07)	41.11 ± 1.02 (39.59-42.46)	58.95 ± 0.82 (58.03-60.03)
C18:2	53.36 ± 0.36 (52.83-53.79)	55.78 ± 3.57 (49.94-59.58)	57.19 ± 1.46 (55.58-58.81)	5.07 ± 1.32 (3.97-7.25)	34.83 ± 9.45 (21.51-42.51)	36.91 ± 0.39 (36.52-37.30)	37.35 ± 1.73 (34.78-39.65)	21.68 ± 0.72 (20.71-22.45)
C18:3	7.22 ± 0.32 (6.89-7.77)	1.59 ± 0.22 (1.25-1.85)	0.59 ± 0.11 (0.50-0.77)	0.88 ± 0.05 (0.79-0.92)	4.24 ± 5.39 (0.40-11.86)	1.95 ± 0.17 (1.78-2.12)	1.64 ± 0.37 (1.11-2.15)	12.03 ± 0.19 (11.76-12.19)

[a] The wt% ± SD is upper number paired with each fatty acid in column and (minimum-maximum) is lower number (in parentheses).

TABLE 2

Fatty Acid Compositions of Coconut Oil as Methyl Esters[a]

Fatty acid	Coconut (three kinds)
C6:0	0.62 ± 0.02 (0.60–0.64)
C8:0	8.05 ± 0.10 (7.97–8.19)
C10:0	5.89 ± 0.19 (5.73–6.16)
C12:0	44.78 ± 1.06 (43.28–45.54)
C14:0	17.69 ± 0.49 (17.31–18.38)
C16:0	10.13 ± 0.17 (9.97–10.37)
C16:1	— —
C18:0	2.87 ± 0.34 (2.39–3.18)
C18:1	6.48 ± 0.31 (6.18–6.91)
C18:2	1.73 ± 0.27 (1.50–2.11)
C18:3	— —

[a]See footnote to Table 1.

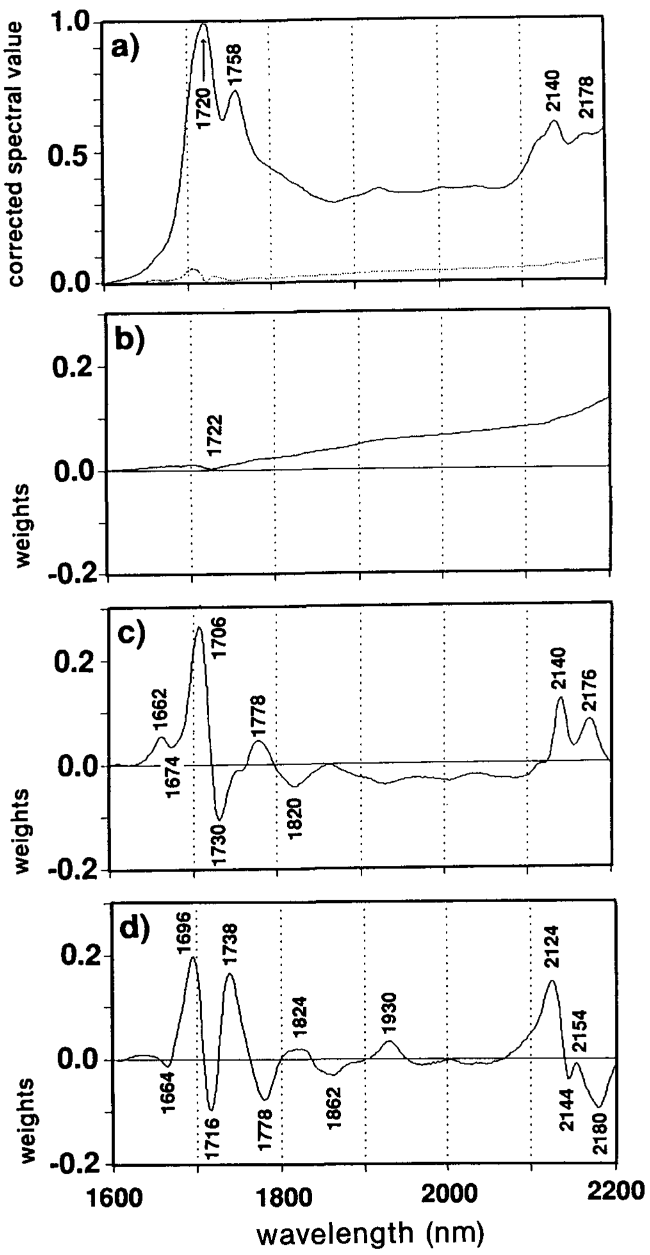

FIG. 1. Results of principal-component analysis of standardized raw near-infrared spectra of oils measured: (a) shows average (solid line) and standard deviation (dotted line) spectra, and (b)–(d) provide the first three eigenvectors of principal-component analysis.

its variance or standard deviation, because PCA orderly extracts variation from the population. No characteristics seemed to be related to fatty acid composition in the first PC. In the second PC (Fig. 1c), peaks at 1662, 2140 and 2176 nm are characteristic absorption wavelengths for unsaturated fatty acid moieties, as previously reported (12). Triacylglycerols also have striking peaks around 1720 and 1760 nm, as shown in the average spectrum (Fig. 1a). As for the former wavelength, the oleic moiety (C18:1) and saturated moieties have a peak maximum at 1725 nm. The higher the degree of unsaturation, the more the peak maximum shifts to a shorter wavelength: to 1717 nm for the linoleic moiety (C18:2) or to 1712 nm for the linolenic moiety (C18:3) (12). There is a deviation between 1717 (or 1712) and 1706 nm. However, the high loading weight at 1706 nm was caused by the *cis* double bonds, especially in C18:2 or C18:3. Because the NIR absorption bands are broad, PCA could detect different degrees of unsaturation in this region more clearly than at the maximal absorption wavelength. On the other hand, the trough at 1820 nm seems to be caused by saturated moieties as previously reported (12). A trough at 1730 nm seemed to be due to C18:1 and saturated moieties. In brief, the second PC represents a higher degree of unsaturation (C18:2 and C18:3 vs. C18:1 and saturated moieties).

In the third PC (Fig. 1d), peaks at 1824 and 2124 nm are characteristic absorption wavelengths for saturated fatty acid moieties, while troughs at 1664, 2144 and 2180 nm are those for unsaturation. These were almost the same wavelengths found in the second PC. However, the signs were different between the second and third PC, i.e., troughs in the second PC changed to peaks in the third PC and *vice versa.* Also, some new wavelengths appeared whose loading weights were high: 1696, 1716 and 2124 nm. The peak at 1696 nm, which corresponds to 1706 nm in the second PC, seemed to be due to C18:3 as previously mentioned. The trough at 1716 nm is due to C18:1, C18:2 or saturated moieties, while the one at 2124 nm reflects saturation. The characteristics of the third PC were as follows: New wavelengths appeared, and the same absorption wavelengths were also found as in the second PC, but the combinations of their signs were different. The third PC reflects the balance of (C18:1 and C18:2) vs. (saturated moieties) vs. (C18:3). The third PC determines not only higher unsaturation but also compositional quality of the unsaturated moieties.

Figure 2 shows the average and standard deviation of the standardized second-derivative NIR spectral data of

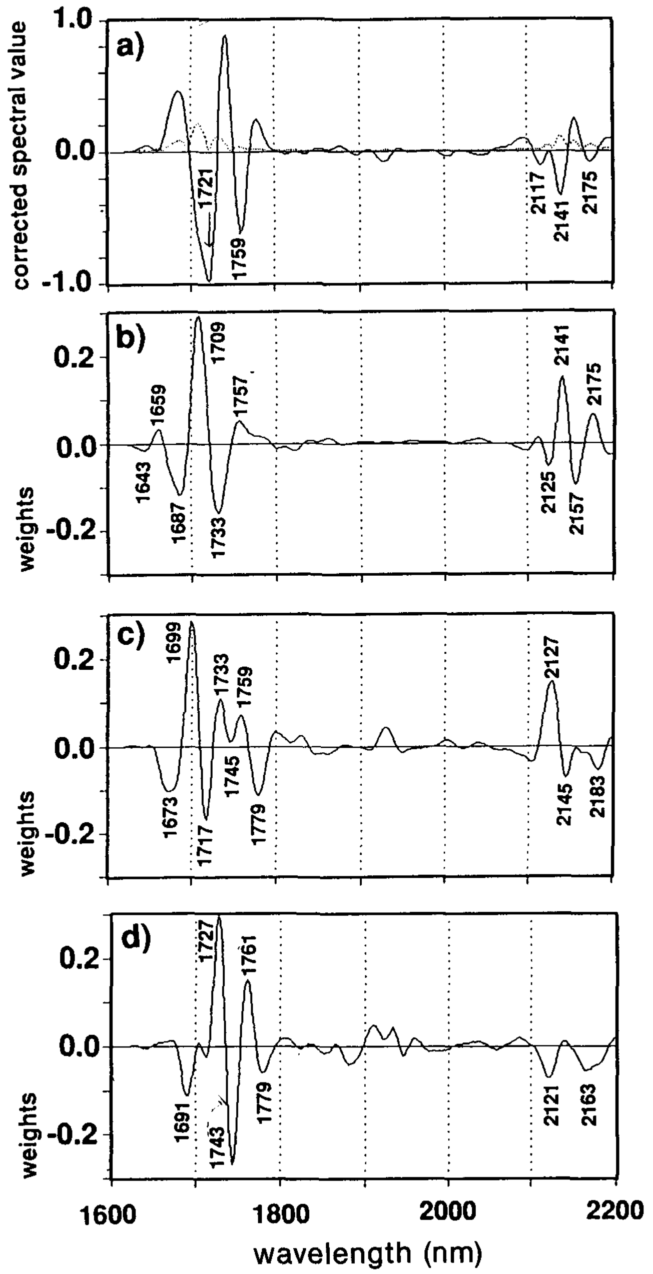

FIG. 2. Results of principal-component analysis of standardized second-derivative near-infrared spectra of oils measured: (a) shows average (solid line) and standard deviation (dotted line) spectra, and (b)–(d) provide the first three eigenvectors of principal-component analysis.

the oil measured and their first three eigenvectors of PC. Their contribution ratio was 95.32, 3.92 and 0.49%, respectively, and their accumulated contribution ratio was 99.73%. The first PC of the second-derivative NIR spectra had similar wavelengths at which the loading weight was as high as shown in the second PC of the raw NIR spectra. In the first eigenvector (Fig. 2b), peaks at 1659, 2141 and 2175 nm were due to absorption wavelengths for unsaturated fatty acid moieties as previously reported (12), and the peak at 1709 nm was also due to C18:2 or C18:3, as already mentioned. The trough at 2125 nm was caused by the saturated acids. The trough at 1730 nm was due to C18:1 and saturated moieties. Troughs at 1687 and 2157 nm seemed to be artifacts, because second-derivative treatment not only separated the overlapping bands but also produced small ghost peaks near the root of the real absorption bands. These wavelengths correspond to peaks in the second-derivative spectrum (Fig. 2a). In the second-derivative spectrum, absorption bands appeared in opposite directions to the raw spectrum after using this mathematical treatment, i.e., troughs appear in the absorption regions, and peaks are artifacts. Another characteristic of the first PC was that its absolute value spectrum was similar with respect to standard deviations, because maximal, minimal and intercept wavelengths correlate with those of the standard deviation. However, the linearly increasing slope, which was found in the first PC of the raw NIR spectra (Fig. 1b), was removed by the mathematical treatment of the second-derivative. PCA first extracted the standard deviation of the whole-range spectrum, but now the standard deviation spectrum reflects the compositional difference from the first.

The second PC (Fig. 2c) of the second-derivative NIR spectrum had wavelengths similar to those shown in the third PC (Fig. 1d) of the raw NIR spectra. The peak at 2127 nm is a characteristic absorption wavelength for saturated fatty acid moieties. On the other hand, troughs at 1673 and 1717 nm represent unsaturated acids. The peak at 1699 nm deviated from the absorption maximum, which means that this wavelength is more effective for distinguishing the differences in the degree of unsaturation of oils. The difference between the first and the second PC was that a peak at 1709 nm shifted to 1699 nm, and this means the second PC signifies more unsaturation. Further, a trough at 1717 nm and a peak at 2127 nm appeared. These seem to be related to C18:1, C18:2 and saturated moieties. The second PC reflects the quality of unsaturation, i.e., the balance of (C18:1 and C18:2) vs. (C18:3) vs. (saturated moieties).

There are no characteristics in the third PC (Fig. 2d). Troughs at 1743 and 1779 nm correspond to wavelengths of actual peaks, which were artifacts produced by the second-derivative treatments, as already mentioned. There are some troughs at 2121 and 2163 nm, but their loading weights were low compared with those at 1743 nm.

Classification by PCA scores. From the first PC of the raw NIR data, tested oils could not be successfully classified because of its nonspecific characteristics. As for raw NIR spectra, the first PC was therefore not used for classification purposes. Figure 3 shows the score plot of the second vs. the third PC for the standardized raw NIR spectra. In the horizontal direction, the order is (coconut) < (olive) < (peanut) < (rice bran) < 0 < (sesame) < (rapeseed, cottonseed) < (corn) < (soybean). This is the order of the degree of unsaturation, specifically the order of the following values: (amount of C18:2) × 2 + (amount of C18:3) × 3, which reflects the predicted results from eigenvectors of the second PC (Fig. 1c). The left side is rich in saturated fatty acid, and as the degree of unsaturation increases, plots moved farther to the right side of the scattering graph. In the vertical direction, the order is (olive) < (peanut) < (rice bran, sesame, rapeseed) < 0 < (corn) < (cottonseed, soybean) < (coconut). The order seemed to be decided by the balance of (C18:1 and C18:2) vs. (saturated moieties) vs. (C18:3).

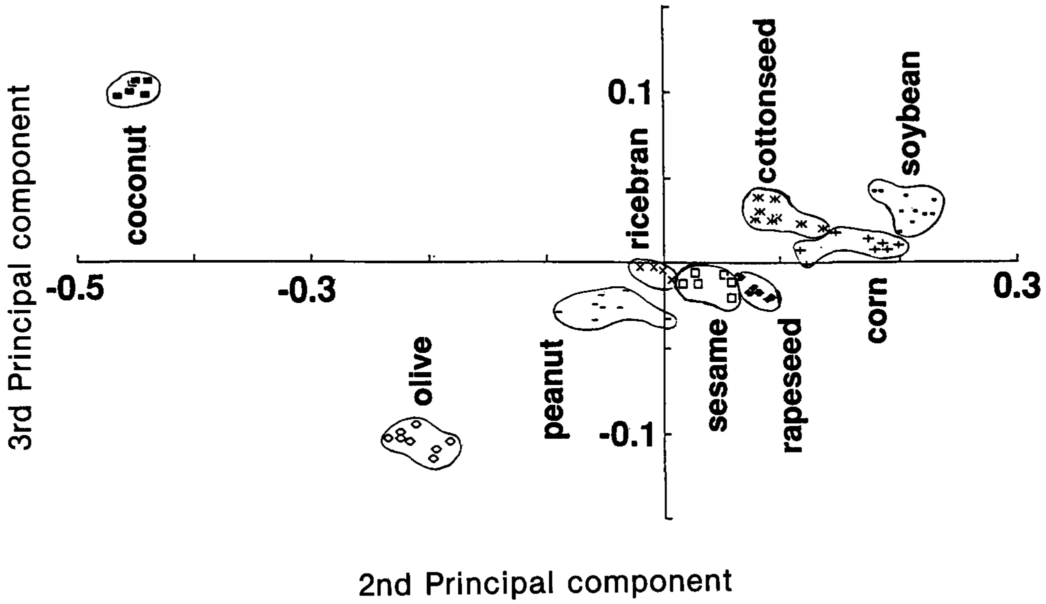

FIG. 3. Score plot of second vs. third principal components for raw near-infrared spectra.

Figure 4 shows the score plot of the first vs. second PCs for the second-derivative NIR spectra. It had the same trends but in an inverse way. On the X-axis, the order is (soybean) < (corn) < (cottonseed) < (rapeseed) < (sesame) < 0 < (rice bran) < (peanut) < (olive) < (coconut). The left side is rich in unsaturated fatty acid, and as the degree of saturation increased, plots moved to the right side. On the Y-axis, the order is (coconut) < (cottonseed, soybean) < (corn) < 0 < (peanut, rice bran, sesame, rapeseed) < (olive). The order seemed to be in decreasing order for the amount of (C18:1 + saturation) as already mentioned. The difference between the raw and second-derivative plots is that when PCA was applied to the second-derivative spectra, the meaningful PC, related to fatty acid composition, was extracted from the beginning, and there is less overlapping between each kind of oil with good resolution for the second-derivatives.

When using nonstandardized spectral values, PCA resulted in score plots, where each oil occupied larger areas and overlapped more with other oils. Standardization makes occupied areas smaller and more clearly separated on the scattering graphs. Standardization is useful to more clearly enhance the characteristics of patterns for NIR spectra of oils. The author has used a variance-covariance matrix instead of a correlation matrix, because in the latter case, the values are divided by the standard

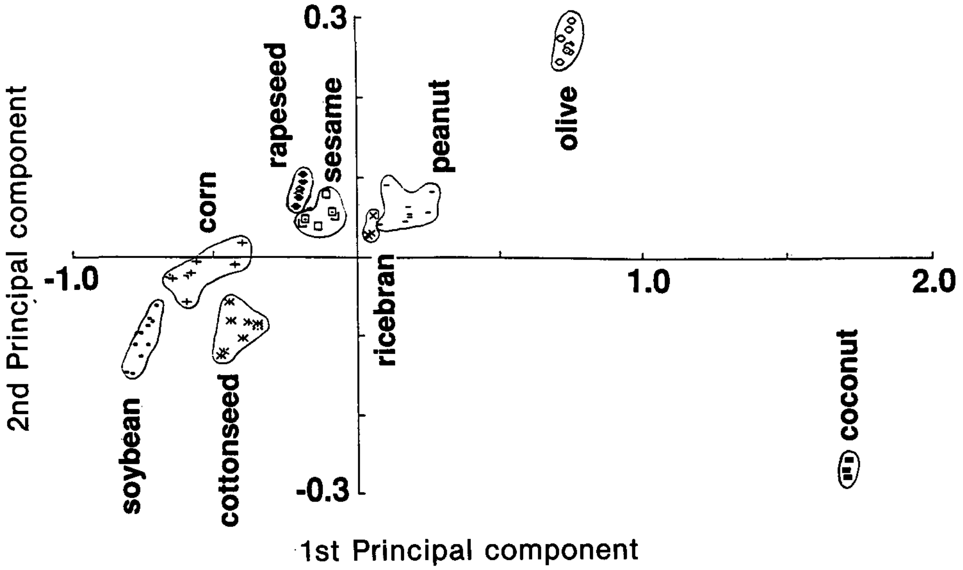

FIG. 4. Score plot of first vs. second principal components for second-derivative near-infrared spectra.

deviation, i.e., normalized, and if the standard deviation is near zero, as in the 1900–2100 nm region, the contribution of these regions was very high while there are no strong characteristic absorption bands.

NIR spectra of oils depend on their fatty acid composition. So, PCA on NIR spectral data depends on variation. PCA score plots belonging to one variety of oil occupied a little bit wider area in the scattering graph than expected, due to the variation in fatty acid compositions shown in Tables 1 and 2. If an oil has a large deviation from the average fatty acid composition of the variety, it may be classified unsuccessfully by PCA. However, in that case, PCA can detect that its fatty acid composition is different from the average. It means that an effective result may be obtained even in that case.

Once the eigenvectors were calculated, PCA scores for other new samples of the same kind of oil may be calculated by summation of the NIR spectral data multiplied by the loading weight, and the oils could be classified rapidly and easily according to these values. As for the NIR measurements, they are easy and there is no need for sample preparation. For use of the NIR method, chemical analysis is always necessary as a reference method. However, by using PCA, classification could be carried out rapidly and easily without a wet chemical method, especially for oils, because NIR spectra contain information about fatty acid composition. Data can be accumulated for other various oils, and this method can then be used for identifying unknown material and for judging adulteration of oils.

ACKNOWLEDGMENTS

The author thanks Dr. Dominique Bertrand of Nantes Research Center, I.N.R.A. (Nantes, France), for his helpful advice on PCA. The author also thanks Dr. Takashi Ide of National Food Research Institute (Tsukuba, Japan) for the gas chromatographic data. The author used the computer system of the Computing Center for Research in Agriculture, Forestry and Fisheries.

REFERENCES

1. Osborne, B.G., and T. Fearn, *Near-Infrared Spectroscopy in Food Analysis*, Longman Scientific & Technical, Harlow, 1986.
2. Williams, P.C., and K. Norris (eds.), *Near-Infrared Technology in the Agricultural and Food Industries*, American Association of Cereal Chemists, St. Paul, 1987.
3. Burns, D.A., and E.W. Ciurczak (eds.), *Handbook of Near-Infrared Analysis*, Marcel Dekker, Inc., New York, 1992.
4. Wold, S., K. Esbensen and P. Geladi, in *Chemometrics Tutorials*, edited by D.L. Massart, R.G. Brereton, R.E. Dessy, P.K. Hopke, C.H. Spiegelman and W. Wegscheider, Elsevier Science Publishers, Amsterdam, 1990, pp. 209–224.
5. Cowe, I.A., and J.W. McNicol, *Appl. Spectros. 39*:256 (1985).
6. Cowe, I.A., J.W. McNicol and D.C. Cuthbertson, *Analyst 113*:269 (1988).
7. Bertrand, D., P. Robert and W. Loisel, *J. Sci. Food Agric. 36*:1120 (1985).
8. Bertrand, D., M. Lila, V. Furtoss and P. Robert, *Ibid. 41*:299 (1987).
9. Downey, G., P. Robert, D. Bertrand and P.M. Kelly, *Appl. Spectrosc. 44*:150 (1990).
10. Robert, P., D. Bertrand and M.F. Devaux, *Anal. Chem. 59*:2187 (1987).
11. Sato, T., S. Kawano and M. Iwamoto, *J. Dairy Sci. 73*:3408 (1990).
12. Sato, T., S. Kawano and M. Iwamoto, *J. Am. Oil Chem. Soc. 68*:827 (1991).
13. Chikuni, K., S. Ozawa, T. Mitsuhashi, M. Mitsumoto, T. Koishikawa, S. Kato and K. Ozutsumi, *Jpn. J. Zootech. Sci. 60*:29 (1989).
14. Tanaka, Y., T. Tarumi and K. Wakimoto (eds.), in *Handbook of Statistical Analysis by Personal Computer, Vol. 2, Multivariate Analysis* (in Japanese), Kyoritsu, Tokyo, 1984, pp. 160–175.

[Received June 9, 1993; accepted December 22, 1993]